UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Mattavelli, S., Van Dessel, P., & De Houwer, J. (2021). Why Does the IAT Reveal a Preference for Stimuli Said to Be Paired With an Unpleasant Sound? Stalking the Unexpected. *Collabra: Psychology, 7*(1). https://doi.org/10.1525/collabra.18733

Social Psychology

Why Does the IAT Reveal a Preference for Stimuli Said to Be Paired With an Unpleasant Sound? Stalking the Unexpected

Simone Mattavelli[1], Pieter Van Dessel[2], Jan De Houwer[2]

[1] Psychology, University of Milan - Bicocca, Milan, Italy, [2] Department of Experimental-Clinical and Health Psychology, Ghent University, Ghent, Belgium

Keywords: iat, automatic evaluations, fear, instructions, evaluative conditioning

https://doi.org/10.1525/collabra.18733

Collabra: Psychology

Vol. 7, Issue 1, 2021

The Implicit Association Test (IAT) is the most widely used measure to assess automatic evaluations. One classic phenomenon that has been well established both using the IAT and self-report measures of liking is evaluative conditioning (EC), which refers to a change in the evaluation of a stimulus due to its pairing with another stimulus. Research has documented that EC can also occur when participants are merely informed about upcoming stimulus pairings. In a recent study, participants reported a more negative evaluation of non-words that were instructed to be followed by an unpleasant sound compared to non-words that would not to be followed by this sound (De Houwer, Mattavelli, & Van Dessel, 2019). Interestingly, however, an unexpected pattern was observed on an IAT, that is, a preference for the stimulus said to be followed by the sound. We report three pre-registered experiments ($N = 650$) in which we manipulated different aspects of the procedure such as the stimuli, instructions, and the measure, but that still revealed the same dissociation. Based on three pilot experiments ($N = 92$), we then conducted a registered report study (Experiment 4) testing whether the unexpected effect depends on how the USs are labelled in the instructions. Although describing the aversive sound as negative eliminated the unexpected IAT effect, the type of instruction only had a weak impact on IAT effects. These results are in line with prior evidence showing that IAT scores are malleable.

Since the advent of implicit social cognition, researchers have increasingly relied on implicit measures[1] to assess whether people like or dislike stimuli. Among the plethora of such measures, the most popular is the Implicit Association Test (IAT, Greenwald et al., 1998). In the IAT, participants rapidly categorize target stimuli (e.g., members of two social groups) and attribute stimuli (e.g., positive and negative words). In one critical block, categories for target and attribute stimuli are combined such that participants need to press a first key for one class of target stimuli and one class of attribute stimuli (e.g., press left for pictures of men and positive words) and a second key for the other class of target and attribute stimuli (e.g., press right for pictures of women and negative words). In another critical block, the response assignments for the target categories are reversed (e.g., press left for pictures of women and positive words; press right for pictures of men and negative words). When attribute categories refer to positive and negative valence, differences in performance between the two blocks are interpreted as evidence for differences in the automatic evaluation of the target categories (e.g., better performance in the men-positive block would reflect a more positive automatic evaluation of men than of women).

One of the phenomena that has often been examined with implicit measures such as the IAT, is evaluative conditioning (EC). EC is defined as the change in the evaluation of a conditioned stimulus (CS) due to its pairing with a valenced unconditioned stimulus (US, De Houwer, 2007). The EC effect has been documented using different measures of liking, including self-report measures and different variants of the IAT (see Hofmann et al., 2010 for a metanalysis). Interestingly, recent advances in EC research have shown that not only actual stimulus pairings, but also instructions about such pairings can generate EC effects (see De Houwer et al., 2020, for a review). The mere act of informing participants that one neutral (CS) and one valenced (US) stimulus will be paired, influences the subsequent evaluation of the CS. Just like the standard EC effect, instruction-based EC has been reliably established both for self-report and IAT measures of evaluation (e.g., De Houwer, 2006; Gast & De Houwer, 2013). For instance, De Houwer (2006) found that after participants were instructed that non-sense words

1 Note that we used the term "implicit" to refer to those measures that probe behaviour under conditions of automaticity (Moors & De Houwer, 2006).

(CSs) would be followed by either positive or negative pictures, they exhibited a preference for CSs said to be paired with positive pictures when self-reporting liking of the CSs and in IAT scores.

Relying on the idea that the same variables might affect both learning via stimulus pairings and learning via instructions, we recently investigated whether an instruction-based procedure could lead to an often observed dissociation between two different instances of learning, that is, expectancy learning and EC (De Houwer et al., 2019). Whilst expectancy learning is heavily affected by the strength of the CS-US contingency (i.e., the difference between the probability of the US when the CS is present and the probability of the US when the CS is absent; e.g., Rescorla, 1966), EC has proven insensitive to it and chiefly depends on mere CS-US co-occurrence (Baeyens et al., 1993; but see Halbeisen & Walther, 2016; Kattner, 2014). In two formally identical studies, we used verbal instructions to manipulate both the statistical contingency and the co-occurrence between two non-words (CSs) and an unpleasant sound (US). The impact of this manipulation was tested on expectancy measures (i.e., self-reported expectancy of the US after presentation of a CS) and on liking measures (i.e., self-reported liking of the CSs measures and an IAT designed to capture automatic evaluation of the CSs). Results showed that the instruction procedure produced the expected changes in expectancies (i.e., expectancy learning) and liking (i.e., EC effects), with a CS_{neg} (i.e., a non-word said to be sometimes followed by the unpleasant US) inducing higher US-expectancy and evaluated less positively than a CS_{noUS} (i.e., a non-word said to be never followed by the sound)[2]. Moreover, expectancy learning but not EC was dependent on instructions about statistical contingency. Importantly, however, the study also revealed an unexpected effect. Whereas EC was robustly observed on the self-report measure, the IAT score revealed a different story, such that the overall IAT score indicated a preference for CS_{neg} over $CS_{noUS.}$ This pattern of results was consistent across the two experiments.

At first glance, our findings are difficult to reconcile with previous research on instruction-based EC, which typically found similar effects on both self-report measures and the IAT (De Houwer, 2006; Gast & De Houwer, 2013; Kurdi & Banaji, 2017). Although it should be noted that none of these studies used the IAT to measure evaluations of stimuli said to be paired with aversive sounds, prior studies testing EC via stimulus-pairings showed that presenting neutral stimuli together with unpleasant sounds affected automatic evaluation of the former class of stimuli in the expected direction (i.e., less liked than stimuli paired with pleasant sounds, Mann et al., 2020; Moran et al., 2015; Moran & Bar-Anan, 2013). For instance, Moran & Bar-Anan (2013) found that target stimuli that co-occurred with an unpleasant scream were less liked than target stimuli that co-occurred with a pleasant melody, as indexed by performances on an IAT and a Sorting Paired Features task (Bar-Anan et al., 2009). Evidence from conditioning studies using other fear-evoking unconditioned stimuli is scarce and offered mixed results when it comes to IAT effects. For instance, Boschen et al. (2007) used the IAT to detect the effects of pairings between one facial image and a mild electrotactile stimulation (as opposed to second facial image presented alone). They found conditioning effects on skin conductance responses but not IAT scores. However, in a fear study with children, Field & Lawson (2003) found that describing unknown animals with negative (versus positive) words produced explicit fear beliefs and also IAT scores in the predicted direction (i.e., preference for animals described as positive).

Taken together, the inconsistency between our findings (De Houwer et al., 2019) and those from prior research on EC based on instructions (e.g., De Houwer, 2006) and stimulus-pairings (e.g., Moran & Bar-Anan, 2013), paves the way for investigating further the IAT effects that we observed in our earlier studies. Given the unexpected nature of the previous results, such an investigation has the potential to reveal new, unexpected insights. For instance, it might reveal new information about why IAT scores sometimes reflect influences unrelated to evaluation (see Meissner et al., 2019 for a discussion). Alternatively, it might present new information about the potential of the IAT to unveil evaluative information that cannot be captured by simple self-report liking. For instance, perhaps people have an automatic preference for stimuli that co-occur with aversive sounds even though they fail to explicitly report this. Once we have identified the reasons for the unexpected results, we will also examine the implications of our findings for theories of (evaluative) conditioning and automatic evaluations.

We report a series of four experiments that employed a procedure similar to De Houwer et al. (2019). Starting from the same basic procedure, we implemented procedural changes to identify potential reasons for the unexpected effect of our instruction manipulation on IAT scores. Across Experiments 1-3 we varied (i) the moment at which the IAT was administered following instructions (Experiment 1); (ii) the nature of the IAT (Experiment 2 and 3); and (iii) the number of stimuli used as US (Experiment 3). Unexpectedly, however, in all three experiments we still observed the same dissociation between IAT scores and self-reported liking. We therefore decided to follow a more systematic approach in which we performed an additional set of pilot studies that started from a very basic set-up that does produce the typical EC instruction effect on IAT scores (i.e., a preference for the CS_{pos}) and then added manipulations until we again found the dissociation. Based on these pilot studies, we conducted a final, fourth pre-registered experi-

2 Note that in De Houwer et al. (2019) we referred to the CS that was told to be sometimes followed by the aversive US as CS+ (i.e., CS signaling the presence of the US) and to the CS that was told to be never followed by the US as CS- (i.e., CS signaling the absence of the US). Here we preferred to use labels that refer to the valence of the US each CS is meant to be paired with. In line with this, we chose CS_{neg} for the CS told to be sometimes followed by an aversive US (Experiments 1-4), CS_{noUS} for the CS told to be never followed by the aversive US (Experiments 1-3), and CS_{pos} for the CS told to be followed by a pleasant US (Experiments 3-4).

ment that tested whether the unexpected effect on the IAT depended on the instructions being either explicit or not about the valence of the US.

Experiments 1-3

Experiment 1 was based on the same method and design adopted in De Houwer et al. (2019). We changed one specific aspect of the procedure, that is, the order of administration of the IAT and the CSs rating phase. In our previous study, CSs ratings were taken right after the onset of each US and before the potential occurrence of the unpleasant sound. The IAT was administrated only after the CS ratings were completed. Crucially however, none of the CSs ratings were followed by the occurrence of the sound. Because the CS rating phase provided a disconfirmation of the instructions (i.e., unlike to what participants had been told, CSs were never followed by USs), the effect of the instructions on the IAT might have been reduced (e.g., if participants no longer believed the instructions) or even reversed (e.g., if participants distrusted the instructions and suspected that the reverse might be true) simply because the IAT was always administrated after the CS rating phase. To test this idea, in Experiment 1, the IAT was presented immediately after the instruction phase. By implementing this change, we expected the impact of instructions on the IAT to be more similar to the effect on self-reported liking (i.e., a preference for CS_{noUS} over CS_{neg}).

Experiment 2 introduced two important new elements. First, IAT instructions made clear that during the IAT participants would not hear any sound (but the sounds would occur in the later phase). This should prevent participants from expecting the occurrence of the unpleasant noise while performing the IAT (which, if participants would find this unexpected absence positive, could drive the unexpected IAT effect). Second, we used a personalized version of the IAT (pIAT, Fazio et al., 1986). In a typical IAT (e.g., with positive, negative, and stimulus category labels), many influences other than liking-related influences can be used to complete the task (see Meissner et al., 2019). Most prominently, participants might draw on specific propositional information that is unrelated to participants' beliefs about the valence of the stimuli (De Houwer, 2014). For instance, in De Houwer et al. (2019), some participants might draw on the propositional information that there is a *positive* relation between the CS_{neg} and a sound (e.g. "CS_{neg} and the unpleasant noise *will* co-occur") and a *negative* relation between the CS_{noUS} and a sound (e.g., "CS_{noUS} and the unpleasant noise *will not* co-occur") when completing the IAT. This might lead to better performance in IAT blocks where participants respond to CS_{neg} with the key assigned to positive words. Circumventing this problem, the pIAT uses "I like" and "I dislike" to capture specific evaluative beliefs about the target stimuli. This should therefore encourage participants to perform the task based on beliefs that relate to their actual liking or disliking the target stimuli.

Experiment 3 further explored the idea that beliefs unrelated to liking might affect IAT performance. One such belief relates to the "figure-ground" effect (see Rothermund & Wentura, 2001). When instructions inform participants that an aversive sound will be sometimes present and sometimes absent after seeing one CS but always absent after seeing another CS, then the presence of the sound becomes more salient (the figure) than the absence (ground). As a result, a CS predicting the US (CS_{neg}) might become more salient. Assuming that the IAT category "I like" is also more salient than the IAT category "I do not like", the salience account would predict faster IAT performance in blocks in which the CS_{neg} and "I like" are assigned to the same key, which is what we observed. To test this idea, we changed two aspects of the procedure. First, we manipulated the instructions referring to a contrast (non-negative) CS ($CS_{contrast}$: CS_{noUS} vs. CS_{pos}) such that for half of the participants they indicated that the CS would never be followed by the unpleasant sound (CS_{noUS}: as before) and for the other participants, we included a pleasant sound and instructions indicated that the CS would always be followed by this pleasant sound (CS_{pos}). In the latter condition, both CS_{neg} and CS_{pos} signaled the presence of a valenced sound. This should reduce differences in CS salience and ultimately prevent a figure-ground effect in the IAT. Second, half of the participants completed a standard IAT, whereas the other participants completed a recoding-free IAT in which the assignment of two categories to the same response switches around from trial to trial (rather than in separate IAT blocks). It has been argued that the recoding-free IAT is less susceptible to the effects of figure-ground asymmetries (Meissner et al., 2019). Hence, if figure-ground asymmetries are responsible for the unexpected effects in the IAT, then the recoding-free IAT should reveal a preference for the $CS_{contrast}$ rather than for the CS_{neg}.

For all the experiments, we pre-registered the materials, sampling plan, exclusion rules, analysis plans, and hypotheses on the Open Science Framework. We report all data exclusions, manipulations, measures, and how we determined our sample sizes. The materials, data and analysis scripts of all experiments are available at osf.io/f6u2n/.

Method

In-line with De Houwer et al. (2019), Experiments 1 and 2 adopted a 2 (CS type: CS_{neg} vs. CS_{noUS}) x 2 (CS-US Statistical Contingency: high vs. low) x 2 (CS-US Co-occurrence: high vs low) mixed design, with the first factor measured within and the other two factors measured between subjects. We also counterbalanced the identity of CS_{neg}, the assignment of each CS to either positive or negative stimuli in the first block of the IAT, and the order in which the two CSs were evaluated (the last two factors were nested).

In Experiment 3, we opted for a 2 (CS Type: CS_{neg} vs. $CS_{contrast}$) x 2 (Statistical Contingency between CS_{neg} and US: high vs. low) x 2 ($CS_{contrast}$ Instructions: absence of negative sound (CS_{noUS}) vs. presence of a positive sound (CS_{pos})) x 2 (IAT type: standard IAT vs. recoding free IAT) mixed design, with the first factor measured within and the other three factors measured between subjects (see Figure 1 for a schematic overview of the procedure adopted in the original and the present studies). All the experiments received approval from the Ethical Committee of the Faculty of Psychology and Educational Sciences at Ghent University.

Participants. Participants were recruited via the Prolific

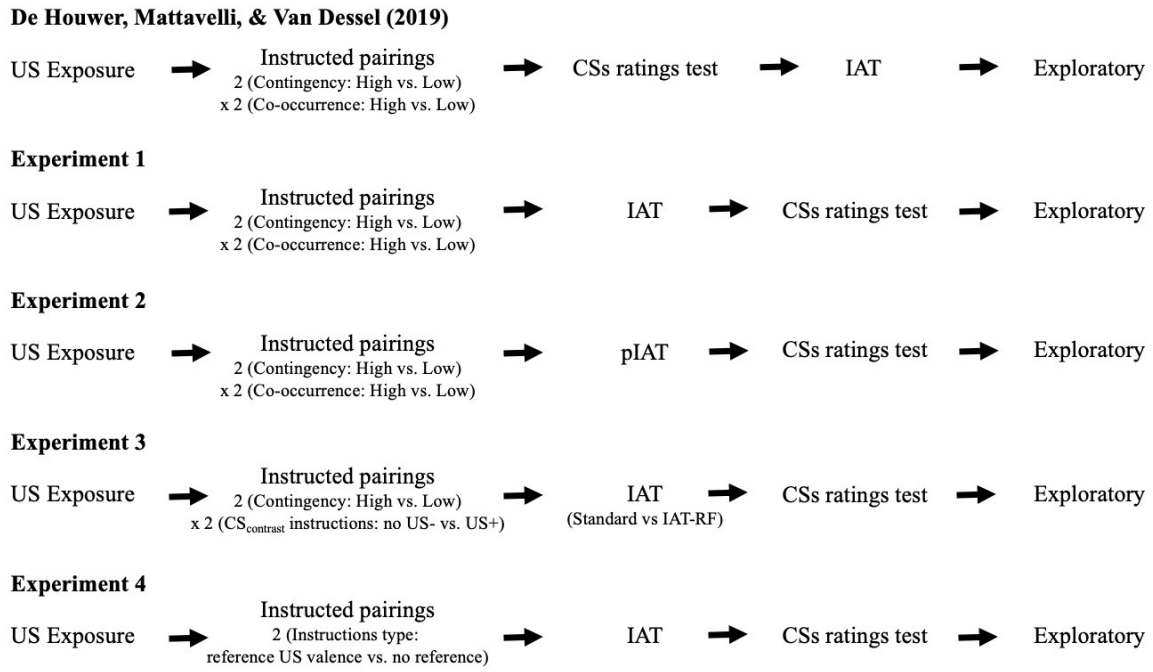

Figure 1. Experimental flow adopted in the original research (De Houwer et al., 2019) and in the four studies.

Academic website (https://prolific.ac/). The target sample size was set via a-priori power analyses conducted on G-Power 3.1. For Experiments 1 and 2, we determined that 200 participants needed to complete our experiment to have high power (> 0.90) to observe a moderate sized effect (d = 0.50) in a planned *t*-test on IAT scores (i.e., a significant preference for the CS_{noUS} over the CS_{neg}), at alpha = 0.05. In Experiment 3, we planned a sample size of 280 participants to have at least 255 participants in the final analyses which provides sufficient power (0.90) to observe a difference of small to moderate effect size (η^2_p = 0.04) in an ANOVA testing the interaction between the manipulated factors of instructions and IAT Type on IAT scores, at alpha = 0.05. In all experiments, we opted for slight oversampling to control for possible exclusions and recruited 228, 230, and 301 participants, respectively (Expriment1: 125 women, *Mage* = 33.71, *SD* = 11.08; Experiment 2: 138 women, *Mage* = 34.30, *SD* = 12.07; Experiment 3: 156 women, *Mage* = 37.12, *SD* = 12.26).

In line with our data analysis plan, participants who (1) failed to complete the whole session or had incomplete data, or (2) did not accurately describe the sound or indicated that they could not hear it, (3) had error rates above 30% across the IAT or above 40% for any one of the four critical IAT blocks, or (4) completed more than 10% of IAT trials faster than 400ms, were excluded from the analyses. Applying these screening criteria led to 198 (Experiment 1), 190 (Experiment 2), and 262 (Experiment 3) participants included in the final analyses.

Stimuli. The same stimuli were used as in De Houwer et al. (2019). The negative US was the sound of a woman's scream. Two nonsense words (MORAG and STRUAN) were used as CSs. It was counterbalanced across participants which word served as CS_{neg} or CS_{noUS} (or CS_{pos}). In Experiment 3, the sound of baby laughter was used as the positive US.

Procedure. ***Experiment 1.*** After providing informed consent and demographic information, participants were instructed to put on their headphones and set the sound volume to 30%. Participants then listened to the unpleasant auditory stimulus and were asked to briefly describe it in an open response format.

Participants received verbal instructions about the number of times in which two different visual stimuli (CS_{neg} and CS_{noUS}) were meant to be followed by the aversive sound they just listened to (US) in a following task. Four different between-subjects instruction conditions resulted from crossing instructed CS-US contingency (high: 80% vs. low: 50%) and instructed CS-US co-occurrence (high: 8 trials vs. low: 4 trials). This instruction phase was followed by a retention test that served to assess participants' recollection of the correct number of events for both words. Otherwise, participants saw the instructions screen again. Both the instructions and the retention test referred to the US stimulus as "the sound", without qualifying it based on its pleasantness.

Participants then completed an IAT. The IAT consisted of three practice blocks and two experimental blocks. It began with 16 practice trials in which the target CSs were sorted under their own categories, followed by 16 practice trials sorting the attribute words. Next, participants completed 64 trials in which CS_{noUS} and positive words shared one response key and CS_{neg} and negative words shared another response key. Participants then practiced sorting the CSs on 16 trials with the reversed response key assignment. Finally, they completed a second set of 64 trials in which the CS_{noUS} shared a response key with negative words and CS_{neg} shared a response key with positive words. On each trial, a valence

word or CS was presented in the center of the screen until the participant pressed one of the two valid keys (i.e., 'E' or 'I'). In case of correct response, the stimulus disappeared and was replaced by the next stimulus, presented 400ms later. If the response was incorrect, a red 'X' replaced the stimulus for 200ms and participants were to correct their mistake to proceed to the next trial.

After the IAT, participants completed a CSs ratings phase that consisted in the individual presentation of the two CSs on screen. Participants were informed that between the offset of the CSs and the potential onset of the aversive sound they would be asked to rate the CSs on two distinct dimensions. The two CSs were presented individually and in counterbalanced order in the center of the screen for 4 seconds, followed by both a valence and a US-expectancy rating scale, administered in counterbalanced order between participants. No sound was displayed after the scales and participants were informed that the condition they have been assigned to involved a single trial for each stimulus. Participants rated the extent to which they liked the two CSs on a 21-category scale (-10 = very negative; 0 = neutral; +10 = very positive). US expectancy ratings were collected on each trial using a 9-point Likert scale presented below the CSs with 3 anchor points: 1 = "not at all", 5 = "uncertain", 9 = "certainly". Finally, participants answered two questions about clarity and credibility of the experimental instructions. In this questionnaire, participants indicated the clarity and believability of the instructions on a 10-point scale and could additionally provide general remarks about the experiment (Raes et al., 2014). A second memory test, identical to that administered after instructions, was repeated at the very end of the session.

Experiment 2. Both the design and the experimental procedure mirrored those used in Experiment 1, except for two main modifications. First, we administered a pIAT. There were two differences between the standard version of the IAT used in Experiment 1 and the pIAT: First, the attribute category labels were the Dutch words for "I like and"I don't like" instead of "Positive" and "Negative". Second, as is typically the case in pIATs, there was no error feedback for the attribute stimuli (because we do not know what stimuli participants like or dislike). Second, instructions preceding the task specified that participants should not expect the occurrence of any sound after presentation of each CS on screen.

Experiment 3. The procedure of the experiment was identical to that of Experiment 1, except for variations in (1) instructions about stimulus pairings (and the presence of an additional auditory US) and (2) the nature of the implicit measure. Specifically, participants assigned to a first instruction condition (with two USs) were exposed to both a pleasant and an unpleasant sound at the very beginning of the session. The pleasant US reproduced the laugh of a baby. The two USs were matched in duration. After hearing the USs, participants received verbal instructions about the number of times the two CSs would be followed by the USs. One CS (CS_{neg}) was said to be followed by the aversive US (either 8/10 times or 8/16 times, depending on Contingency), while the other CS ($CS_{contrast}$) was said to be either never followed by the unpleasant US (CS_{noUS}) or always followed by an alternative US selected for being pleasant (CS_{pos}). The retention test for participants assigned to this condition varied according to the manipulation and so did the options of answer. Participants assigned to a second instruction condition (with one US) had the same instructions as in Experiments 1 and 2. Note that, like in Experiment 1-2, there was no explicit mention to the pleasantness of the USs. The instructions and the retention test referred to the US stimuli as "the first sound" and "the second sound".

Participants assigned to a first measurement condition completed a recoding free IAT (IAT-RF, Rothermund et al., 2006). The IAT-RF consisted of a first block of two blocks with 16 trials in which participants sorted CS_{neg} vs. $CS_{contrast}$ stimuli and positive vs. negative stimuli, respectively, a second block with 32 combined practice trials, and a final block with 128 test trials. In contrast to the typical IAT, compatible and incompatible assignments of categories to responses varied randomly between trials within a single experimental block. Each trial began with response assignments indicated as category labels in the respective corners of the screen. Due to the random switching between compatible and incompatible response, participants were forced to categorize all presented stimuli on the basis of their task-relevant category membership. Participants assigned to the second measurement condition completed the typical IAT. Finally, along with the exploratory questions, participants provided their evaluation of the sound(s) on an 11-point scale ranging from -5 to +5 (extremely unpleasant/extremely pleasant).

Results

We present the results for the effect of CS type on both self-reported liking and IAT scores by reporting the results of a paired-sample t-test (comparing self-report CS ratings) and a one-sample t-test (comparing the IAT score to zero), respectively. For Experiments 1-2, we also report results of an ANOVA examining the impact of both Contingency and Co-occurrence on both self-report ratings and IAT score. For Experiment 3, we report results of an ANOVA examining the impact of Contingency (there was no co-occurrence manipulation), IAT type and US instructions.[3] Because not of crucial importance for the purpose of the present project, we do not report results for US expectancy ratings in the main text (results on US expectancy for all the experiments are available at osf.io/f6u2n/).

Experiment 1. ***Self-report ratings.*** A paired-sample *t*-test

3 All the analyses on the IAT data were re-run including block order as additional factor. We did not find significant main effects (*p*s > .30) or interactions (*p*s >.21) in Experiments 1 and 2. In Experiment 3, there was a significant main effect of block order, $F(1,246) = 16.27$, $p < .001$, and an interaction with IAT type, $F(1,246) = 16.93$, $p < .001$. For all the experiments, inclusion of this factor in the analyses did not change the impact of any of the crucial variables on the IAT score.

indicated a typical EC instruction effect, $t(197) = -3.79$, $p < .001$, $d = .27$. Participants indicated less liking for the CS_{neg} ($M = -0.76$, $SD = 4.41$) than for the CS_{noUS} ($M = .73$, $SD = 4.56$). A mixed ANOVA with CS Type as within-subjects factors and Contingency and Co-occurrence as between-subjects factors showed only the main effect of CS Type, $F(1,194) = 13.47$, $p < .001$, but no significant interaction between CS Type and Contingency, CS Type and Co-occurrence, or the three considered together in a three-way interaction, $ps > .27$.

IAT scores. IAT scores were calculated using the D2-algorithm (Greenwald et al., 2003) so that a positive score indicates a preference for the CS_{neg} over the CS_{noUS}. Split-half reliability was $r(196) = .69$. Contrary to our hypotheses, we observed an overall preference for CS_{neg} ($M = .09$, $SD = .44$), $t(197) = 2.97$, $p = .003$, $d = .21$.[4] The factorial ANOVA included Contingency and Co-occurrence as between-subjects factors (but not CS Type because the IAT score is a relative score so the main effect of CS Type is reflected in the Intercept). This ANOVA revealed neither main effects of Contingency or Co-occurrence, nor an interaction between the two, $ps > .14$. We also did not observe a significant correlation between the IAT score and self-reported liking difference scores, $r = .12$, $p = .09$, $BF_{01} = 1.55$.

Experiment 2. ***Self-report ratings.*** A paired-sample *t*-test indicated an EC effect, $t(189) = -3.78$, $p < .001$, $d = .27$. Participants indicated less liking for the CS_{neg} ($M = -1.02$, $SD = 4.69$) than for the CS_{noUS} ($M = .54$, $SD = 4.76$). A mixed ANOVA showed an effect of CS Type, $F = 14.22$, $p < .001$, but no significant interactions with either Contingency or Co-occurrence, or both, $ps > .31$.

IAT scores. Split-half reliability was $r(188) = .62$. An overall preference for the CS_{neg} emerged, $M = .08$, $SD = .41$, $t(189) = 2.64$, $p = .009$, $d = .19$. Neither main effects of Contingency or Co-occurrence nor an interaction between the two was found, $ps > .30$. Again, we observed no significant correlation with self-report liking difference scores, $r = .03$, $p = .72$, $BF_{01} = 5.58$.

Experiment 3. ***Self-report ratings.*** The mixed ANOVA with CS Type as within-subjects factor and Contingency, IAT Type and $CS_{contrast}$ instructions revealed an effect of CS Type, $t(261) = -4.83$, $p < .001$, $d = .30$, indicating less liking for the CS_{neg} ($M = -.73$, $SD = 4.82$) than for the $CS_{contrast}$ ($M = 1.05$, $SD = 4.76$). No other main effects or interaction effects were observed, $ps > .12$.

IAT scores. Split-half reliability for both the standard and the IAT-RF were $r(136) = .81$ and $r(122) = .42$, respectively. We found an overall preference for the CS_{neg}, $M = .06$, $SD = .34$, $t(261) = 2.60$, $p = .010$, $d = .16$. The 2 (Contingency: High vs. Low) × 2 (IAT Type: standard IAT vs IAT-RF) x 2 ($CS_{contrast}$ Instructions: absence of negative sound vs. presence of a positive sound) ANOVA did not reveal any main effects or interactions, $ps > .06$. There was a significant correlation between IAT scores and self-report liking, $r = .16$, $p = .012$, $BF_{10} = 3.22$.

US evaluation. Participants who were exposed only to the unpleasant sound rated it as negative ($M = -3.17$), $t(127) = -13.32$, $p < .001$. Participants exposed to both the USs evaluated the pleasant US more positively ($M = 1.96$) than the unpleasant ($M = -2.57$), $t(133) = -12.65$, $p < .001$. For participants exposed to a single sound, self-report liking towards the CS_{neg} was not correlated with the evaluation of the US, $r = .08$, $p = .37$. Instead, for participants exposed to both USs, self-reported CS_{neg} liking significantly correlated with evaluation of the unpleasant US, $r = .23$, $p = .009$, and CS_{pos} liking correlated with that of the pleasant US, $r = .29$, $p = .001$. No significant correlation between IAT scores and US evaluations was found.

Combined Results

Data from the three experiments were combined ($N = 650$) to investigate the overall effect of the instruction-based conditioning procedure on both self-report ratings and IAT score. Because we were not interested in any other effect at this stage, none of the additional variables manipulated across the three studies (i.e., contingency, co-occurrence, IAT Type and $CS_{contrast}$ instructions) were entered in this analysis. Bayes factors were reported for each analysis to quantify the evidence for the alternative over the null hypothesis (BF_{10}).

Self-report ratings. A paired-sample *t*-test revealed a preference for the CS told to be never paired with the unpleasant sound, $t(649) = -7.13$, $p < .001$, $d = .28$, $BF_{10} > 1000$.

IAT scores. A one-sample *t*-test revealed an overall preference for CS_{neg}, $t(649) = 4.75$, $p < .001$, $d = .19$, $BF_{10} = 2738.16$. The correlation between the IAT score and self-report ratings was significant, although weak, $r = .10$, $p = .008$, $BF_{10} = 2.99$.

Discussion

Across three experiments, we consistently replicated the results of De Houwer et al. (2019). Most importantly, we observed that on a self-report measure of liking, a CS said to be paired with an aversive auditory sound was evaluated less positively than a CS said to be never followed by the US (Experiment 1-2) or always followed by a pleasant auditory sound (Experiment 3). In contrast, this evaluative pattern never emerged when looking at the effect on the IAT. Experiments 1-3 explored several explanations of this unexpected effect in terms of measurement order, the type of IAT measure, and whether instructions referred to either the consistent absence of the negative US or to the presence of a positive US after the contrast CS. Surprisingly, the difference in results on self-reported liking versus IAT scores was observed in all of the conditions. Although the three studies add robust evidence for the unexpected findings of De Houwer et al. (2019), it remains unclear what might explain the inconsistency with previous results on instruction-based EC (De Houwer, 2006; Gast & De Houwer, 2013).

4 Although interpreting IAT score as either positive or negative as compared to the zero point might be see as problematic (Blanton & Jaccard, 2006), counterbalancing the identity of the CS assigned to the aversive US makes such comparison a valid indicator of liking in the IAT.

In a final effort to try and explain these findings, we used an approach in which we wanted to first establish the typical instruction-based EC effect with the current stimuli and then move more closely to the current procedure until we could identify the procedural aspect that eliminates the typical effect. To this end, we conducted a number of pilot studies (materials, data and analyses code for the pilot studies are available at osf.io/f6u2n/). In a first pilot study, we simply instructed participants that one CS nonword would always be followed by an unpleasant sound, while the other CS nonword would be followed by a pleasant sound. Importantly, in this study, instructions made the valence of the USs explicit (i.e., a sound was labelled as pleasant or unpleasant), and participants were never exposed to the USs. We adopted this procedure because it mirrored what has been done in previous studies testing the impact of EC-instructions on IAT scores (e.g., De Houwer, 2006). As expected, a sample of 39 Prolific Academic participants (26 women, *Mage* = 39.03, *SD* = 13.21) exhibited a preference for the CS_{pos} over CS_{neg} on the IAT, M = -.17, SD = .43, $t(38)$ = -2.50, p = .017, d = .40.

The first pilot study showed the expected effect of instructed CS-US pairings on the IAT score. However, there were multiple procedural differences between this pilot study and the previous experiments that might have accounted for the discrepancy in the results. Among such differences, one refers to the contingency between the CSs and the USs. In fact, whereas in the first pilot study instructions indicated perfect contingencies (i.e., CSs always followed by either a negative or a positive sound), in the previous experiment involving both a positive and a negative US (Experiment 3) the contingency between the CS_{neg} and the US was either high (8/10 trials) or low (8/16 trials), but never perfect. Therefore, in a second pilot study, 20 Prolific Academic participants (10 women, *Mage* = 32.80, *SD* = 14.17) underwent the exact same procedure but this time both CS_{neg} and CS_{pos} were instructed to be either followed by the unpleasant (CS_{neg}) or by the pleasant sound (CS_{pos}) in only 8 out of 10 trials. This contingency manipulation did not appear to be the critical factor producing the dissociation as participants again revealed a similar sized preference for the CS_{pos} over the CS_{neg} on the IAT (M = -.21, SD = .43), $t(19)$ = -1.96, p = .065, d = .44. Data from the two pilot studies were combined (N = 59) to estimate evidence for the overall impact of the instruction-based conditioning procedure on self-report and IAT scores, also with Bayes factors. A paired-sample t-test revealed a preference for CS_{pos} on self-report ratings, $t(58)$ = -3.54, p < .001, d = .46, BF_{10} = 32.73. A one-sample t-test revealed a preference for CS_{pos} on IAT scores, consistent with self-report liking, but contrasting with the results of Experiments 1-3, $t(58)$ = -3.20, p = .002, d = .42, BF_{10} = 13.34.

Having ruled-out an explanation in terms of CS-US contingency, we turned our attention to another key difference between the pilot studies and prior experiments, that is, the way the USs were described in the instructions. In Experiments 1-3, participants were first exposed to one sound (or two sounds, Experiment 3) and then received instructions that referred to the presence or absence of one sound after the presence of either CS. However, those instructions did not refer to the valence of the sound(s), which was supposed to have been learned in the initial exposure of the sound(s). Instead, there was no exposure to the sounds in the pilot studies and participants were informed about the valence of the US via instructions. A third pilot study changed this aspect of the procedure. Thirty-three participants (21 women, *Mage* = 35.73, *SD* = 12.54) were exposed to the same USs used in Experiment 3 at the very beginning of the study and then saw the contingency instructions. Importantly, and in line with the procedure used in Experiment 3, these instructions now referred to the USs as either a "first" or a "second" sound rather than "positive" or "negative" sounds. Crucially, this time the IAT score did not indicate a preference for the CS_{pos} (M = -.01, SD = .54), $t(32)$ = -.06, p = .95, d = .01.

Taken together, the results of these pilot studies support a new explanation for the unexpected effects of EC instructions on IAT scores that were observed in Experiments 1-3. In those studies, instructions about CS-US relations did not specify the valence of the US, whereas instructions in studies that did find typical instruction effects on IAT scores always informed participants about the valence of the stimuli that would be paired with the US (e.g., De Houwer, 2006; Gregg et al., 2006; Kurdi & Banaji, 2017). For instance, De Houwer (2006) instructed participants that either a "positive" or a "negative" stimulus would follow the onset of the target non-words and Kurdi & Banaji (2017) told participants that members of two alternative social groups were about to be paired with "positive" or "negative" things.[5] In studies that use stimuli for which valence is very salient (e.g., images of death people), stimulus presentation may evoke labeling in valenced terms and presenting valenced labels in the instructions might not be crucial. However, as suggested by previous inconsistencies in finding effects on the IAT following conditioning with fearful stimuli (e.g., Boschen et al., 2007), reference to the valence of the US stimuli might be very important to qualify the nature of this type of stimuli. Without such reference, fearful stimuli might not be categorized by participants in negative terms (e.g., the "unpleasant" stimulus) but in neutral terms or even positive terms (e.g., the "crucial", "exciting", or "important" stimulus). If we assume that such labelling of the USs also influences the way in which the CSs are labelled during the IAT (e.g., "nonword that was paired with the unpleasant sound" vs. "nonword that was paired with the crucial sound") this could produce an impact of US labelling on IAT performance such that the presence or absence of a

5 An anonymous reviewer pointed out that prior unpublished research found evidence of instruction-based EC on the IAT with no reference to the valence of the US in the instructions. In this research the authors used a modified version of the procedure used in Kurdi & Banaji (2017). Instructions were modified to inform participants that one CS was always paired with things like flowers, a heart, ice cream, the sun, and a beach, whilst another CS was said to be always paired with things like a frowny face, a fleeing man, a snake, a terrorist, and a bug. Note that, although not mentioned explicitly, the use of multiple positive and negative USs might have made the valence of the stimuli salient and therefore influenced the subsequent categorization (or labelling) of the US stimuli.

valenced label determines whether the IAT score indicates more positive or more negative evaluation of the US.

For instance, if participants in our main experiments labelled CS_{neg} as "the nonword that was paired with the unpleasant sound", this might have facilitated responding to that CS in the same way as responding to unpleasant attribute stimuli, leading to IAT scores indicative of a negative evaluation of the CS_{neg}. In contrast, if participants labelled CS_{neg} in a different way, such as "the nonword that was paired with the crucial sound", this might have facilitated responding to the CS_{neg} with the same key as responding to pleasant attribute stimuli, leading to IAT scores indicative of a positive evaluation of the CS_{neg}. Whereas labelling CS_{neg} as "the nonword that was paired with the crucial sound" seems likely when the unpleasant sound is the only US told to be paired with either CS (Experiment 1-2), it might be less likely when both a pleasant and an unpleasant US are used (Experiment 3). However, a fearful sound is not only more negative, but also more arousing than pleasant sound. Past research has shown that the conditioning of affective responses can occur simultaneously along one evaluative dimension (i.e., valence) and along the dimension of arousal (Gawronski & Mitchell, 2014). Therefore, it is still possible that CS_{neg} was conceived as "the nonword that was paired with the arousing sound". Although these considerations are somewhat speculative at this stage, they are in line with the fact that IAT performance depends heavily on the nature of the category labels that are used in the IAT, with identical stimuli producing different effects depending on the labels that are used to categorize them. For instance, in Mitchell et al. (2003) the automatic evaluations of the very same black vs. white social stimuli changed based on the reference category used to sort stimuli throughout the IAT (e.g., the name "Michael Jordan" to be categorized as either a Black person or an athlete). Whereas differences in automatic evaluations observed in that study emerged from a direct manipulation of the category labels, here we propose that even in the absence of any direct manipulation, participants might spontaneously use specific labels when categorizing target stimuli during the IAT, and that such a labelling might ultimately affect their automatic evaluations.

To test the idea that the unexpected IAT effect depends on whether instructions inform participants about the valence of the US, we conducted a well-powered fourth experiment. In this study, we first exposed all participants to both the sounds, and then we directly compared a condition in which instructions overtly referred to the valence of US sounds with effects of an alternative procedure in which instructions did not refer to their valence. This experiment was tightly matched to the procedure of Experiment 3 in which participants were exposed to two sounds, were given instructions that one CS would be followed by the first sound 8/10 times and another CS would be always followed by the second sound, and received a typical IAT. Importantly, however, for one group of participants the instructions referred to 'an unpleasant sound' and 'a pleasant sound'. Therefore, participants in both the experimental and the control condition were pre-exposed to the pleasant and the unpleasant sound, but only for the former the valence of such sound was explicit in the contingency instructions.

Experiment 4

Experiment 4 was run as a registered report. The design and analysis and sampling plan were preregistered at OSF and approved at Collabra:Psychology (https://osf.io/5necg). There were no deviations to report. All data and analysis code are available in the OSF repository at osf.io/8kguz/.

Method

Participants and Design

Experiment 4 had a 2 (CS type: CS_{neg} vs. CS_{pos}) x 2 (Instructions type: explicit reference vs. no reference to US valence) mixed design, with the first factor manipulated within and the second between participants. We also counterbalanced across participants (i) the identity of the CS to be related with either US, and (ii) the order of presentation of the two CSs in the evaluative ratings and the order in which the two critical blocks of the IAT were administered (the two factors will be nested into a single one).

As in the previous experiments, participants were recruited via Prolific Academic. In line with our sampling plan, we stopped data collection when 200 participants completed the entire session (84 females, 79 males, 37 missing information, M_{age} = 33.13, SD_{age} = 12.76). Considering our inclusion criteria, we expected a screened-out rate of approximately 10% of the entire sample so this sample size would allow us to have a good power (0.95) to observe a moderate difference between the two conditions (d = 0.50) in a one-tailed t-test at alpha = 0.05 (required N=176)[6].

We adopted the same exclusion criteria as in the previous and excluded the data of participants who (1) failed to complete the whole session or showed incomplete data (4 participants), (2) did not accurately describe the sound or indicated that they could not hear it (1 participant), (3) had error rates above 30% across the IAT or above 40% for any one of the four critical IAT blocks (4 participants), or (4) com-

6 Because the critical test is the difference between IAT scores in the two instructions conditions, we estimated the effect size considering the effects observed in the previous experiments and pilot studies. The IAT scores in the studies where we used instruction with no reference to the valence of the US (Experiments 1-3 and Pilot 3) are overall in favor of CS_{neg} (Exp1: M = .09, SD = .44; Exp2: M = .08, SD = .41; Exp3: M = .06, SD = .34; Pilot3: M = -.01, SD = .54). Instead, in Pilot 1-2, conducted using the alternative instructions (those with reference to the valence of the US), the IAT score was in the opposite direction, therefore revealing a preference for CS_{pos} over CS_{neg} (Pilot1: M = -.17, SD = .43; Pilot2: M = -.21, SD = .43). A comparison between the average IAT score in the first instruction condition (M = .08) and that observed in the alternative condition (M = -.19), divided by a pooled standard deviation (.43), would lead to d = 0.58. Based on this, we opted for a slightly more conservative effect size d = 0.50 to estimate the required sample size.

pleted more than 10% of IAT trials faster than 400ms (2 participants). Analyses were performed on the data of the remaining 190 participants.

Procedure.

The stimuli, measures, and instructions were identical to Experiment 3 with few exceptions. After participants provided informed consent and answer demographic questions, they listened to the two auditory stimuli (US+ and US-) and briefly described them.

Next, participants were assigned to two instructions conditions. In the explicit instructions condition, participants were informed that in 8 out of 10 trials one CS was followed by the unpleasant sound they listened to, and that the other CS was always followed by the pleasant sound they listened to. For participants in the control condition, instructions referred to "the first[second] sound", without any explicit reference to their pleasantness. Participants then underwent the same retention test administered in Experiment 3. In this test, they were asked to indicate the correct number of times in which, based on instructions, the relevant US would occur upon presentation of each CS. Importantly, neither the questions nor the response options presented in the control condition made explicit reference to the pleasantness of the USs, instead they were simply defined as the "first" and the "second" sound. In case of incorrect responding, participants were exposed to the initial instructions again. Next, participants completed the IAT, followed by the same CSs rating phase used in the previous studies, in which both self-reported liking towards the CSs (i.e., "How much do you like this stimulus?"; response scale: -10 = very negative; 0 = neutral; +10 = very positive) and US expectancy (i.e., No reference to US valence: "To what extent do you expect that the first [second] sound you listened to will occur?"*;* Explicit reference to US valence: "To what extent do you expect that that the pleasant [unpleasant]sound you listened to will occur?"; response scale: 1 = not at all, 5 = uncertain, 9 = certainly) were measured. Participants answered exploratory questions, including the same questions about clarity and believability as in our previous experiments. We also measured self-report evaluation of both the sounds (i.e., "How pleasant was the first(second) sound you listened to at the beginning of the study?") on a 11-point scale (-5 = very unpleasant; +5 = very pleasant). The same memory question administered in the retention test was repeated at the end of the session.

Hypotheses

We expected to find a difference in the EC effect on IAT scores as a function of the type of instructions. Namely, the IAT score should reveal a preference for CS_{pos} over CS_{neg} for participants in the explicit valence instructions condition. Participants assigned to the control condition should show a preference for the CS_{neg}, as was the case in Experiment 3. We also expected to replicate prior findings on self-reported liking: in both the instruction conditions, participants should exhibit a preference for CS_{pos} over CS_{neg}.

Results

Preregistered Analyses

Self-report ratings. A paired-sample t-test showed that the CS_{neg} was evaluated more negatively (M = 0.20, SD = 3.95), than the CS_{pos} (M = 1.30, SD = 3.54), $t(189) = -2.93$, p = .002, d = .21. The mixed-ANOVA, with Instructions Type (explicit reference vs. no reference to US valence) as between-subjects factor and CS Type (CS_{neg} vs. CS_{pos}) as within-subjects factor, showed a main effect of CS Type, $F(1,188) = 8.35$, p = .004, η_p^2 = .04, and a significant interaction, $F(1,188) = 5.86$, p = .02, η_p^2 = .03. In accordance with our hypotheses, a preference for CS_{pos} (M = 1.51, SD = 3.81) over CS_{neg} (M = -0.45, SD = 3.97) was evident when instructions were explicit about the valence of the sounds, $t(97) = -3.25$, p = .002, d = 0.33, BF_{10} = 38.96. However, in contrast with our hypotheses and with the results of Experiments 1-3, we did not observe this preference for CS_{pos} (M = 1.08, SD = 3.24) over CS_{neg} (M = 0.90, SD = 3.82) when instructions did not refer to the valence of the US, $t(91) = -0.42$, p = .67, d = 0.04, BF_{10} = 0.48.

IAT score. Split-half reliability was $r(188)$ = .74. The overall IAT score revealed no preference for either CS (M = 0.01, SD = 0.50), $t(189) = 0.15$, p = .88, d = 0.01. The crucial, expected, difference between the two IAT scores was significant in the planned one-tailed t-test, $t(188) = 1.72$, p = .043, d = 0.25, BF_{10} = 1.49. In contrast to our hypothesis and the results of Experiment 3, we did not observe a significantly positive IAT score (i.e., preference for CS_{neg} over CS_{pos}) in the control condition (i.e., when instructions did not mention the valence of the USs, M = 0.07, SD = .46), $t(91) = 1.44$, p = .076, d = 0.15, BF_{10} = 1.34. We also did not observe a significantly negative IAT score (i.e., preference for CS_{pos} over CS_{neg}) in the explicit instructions condition (i.e., when instruction did mention the valence of the USs, M = -0.05, SD = .53), $t(97) = -1.02$, p = .16, d = 0.10, BF_{10} = 0.81. The IAT score and the self-report liking ratings were positively correlated, $r(188)$ = .19, p = .008.

US expectancy. Expectancy of the unpleasant US was higher for CS_{neg} (M = 6.23, SD = 3.32), as compared to CS_{pos} (M = 3.07, SD = 2.59), $t(189) = 11.78$, $p < .001$, d = .85. Similarly, expectancy of the pleasant US was higher for CS_{pos} (M = 6.61, SD = 2.72), as compared to CS_{neg} (M = 2.82, SD = 2.30), $t(189) = -13.06$, $p < .001$, d = .95. Two mixed ANOVAS that included the factors CS Type and Instruction Type showed no significant interaction effect on either US negative expectancy, $F(1,188) = .06$, p = .80, or US positive expectancy, $F(1,188) = .19$, p = .66.

Non-preregistered Analyses

Comparison between self-report ratings and IAT score. A mixed ANOVA with Measurement type (IAT vs. self-report) as within-subjects factor and IAT order, CS type and Instructions type as fixed factors showed a significant effect of Measurement Type, $F(1,182) = 6.23$, p = .01, η_p^2 = .03, indicating stronger effect when preferences were measured via self-report. We also observed a main effect of Instructions type, $F(1,182) = 8.21$, p = .005, η_p^2 = .01, but no interaction effect of Instructions Type and Measurement, $F(1,182) = .71$,

$p = .40$, $\eta_p^2 = .004$.

US evaluation. The evaluation of the unpleasant US was more negative ($M = -3.72$, $SD = 1.86$) than that of the pleasant US ($M = 2.19$, $SD = 2.24$), $t(189) = -25.55$, $p < .001$, $d = 1.85$. When US evaluations were submitted to a mixed ANOVA with Instructions type as fixed factor, we did not find a significant interaction, $F(1,188) = 1.72$, $p = .19$. Instruction Type also did not show any significant impact on evaluation ratings of either US when considered in isolation (unpleasant US: $F(1,188) = .23$, $p = .63$; pleasant US: $F(1,188) = 2.14$, $p = .15$).

Correlation analyses. For both the self-report and the IAT scores, we inspected the correlations with the other relevant scores (i.e., expectancy of the unpleasant US, expectancy of the pleasant US, US evaluation). A differential liking score (CS_{pos}-CS_{neg}) correlated negatively with the differential expectancy of the unpleasant US (CS_{pos}-CS_{neg}), $r(188) = -.30$, $p < .001$, and correlated positively with the differential expectancy of the pleasant US (CS_{pos}-CS_{neg}), $r(188) = .30$, $p < .001$, and with differential US evaluation ratings (US_{pos}-US_{neg}), $r(188) = .23$, $p = .001$.

We did not observe a significant correlation between IAT scores and ratings of the expectancy of the unpleasant US, $r(188) = .11$, $p = .12$, the expectancy of the pleasant US, $r(188) = -.08$, $p = .243$, and US evaluation, $r(188) = -.02$, $p = .793$.

Meta-analysis. Findings from Experiment 4 diverged from previous experiments with respect to the impact of EC instructions (in the condition that did not refer to the valence of the USs) on self-reported ratings. Therefore, we performed a meta-analysis of the results of the four experiments that examined the impact of EC instructions when instructions did not refer to the valence of the sound(s). Data from the four experiments were combined, with the exception of participants that in Experiment 4 were assigned to the condition in which instructions referred to the pleasantness of the USs. This led to a final sample of 742 participants. Self-report ratings showed a significant preference for the CS_{pos} over the CS_{neg}, $t(741) = -4.84$, $p < .001$, $d = 0.18$, $BF_{10} > 1000$. The IAT score revealed the opposite pattern, with a significant preference for the CS_{neg}, $t(741) = 4.93$, $p < .001$, $d = 0.18$, $BF_{10} > 1000$. Neither Self-report ratings nor IAT scores were qualified by type of experiment, $ps > .39$.

General Discussion

The Implicit Association Test (IAT) is typically used as a measure of automatic evaluations (i.e., evaluations that occur under automaticity conditions). Research has shown that evaluative conditioning procedures (i.e., the pairing of a neutral stimulus, US, with a valenced stimulus, CS) typically lead to changes in IAT scores reflecting a preference for USs paired with positive CSs over USs paired with negative CSs (Hofmann et al., 2010). Similar effects have been observed when participants are merely instructed about upcoming CS-US pairings (e.g., De Houwer, 2006; Gast & De Houwer, 2013; Kurdi & Banaji, 2017).

In a recent study, De Houwer et al. (2019) found an unexpected IAT effect of EC instructions that referred to upcoming pairings of an unpleasant sound (US) with non-words CSs. Specifically, IAT scores reflected a preference for a CS told to be followed by the unpleasant US (CS_{neg}) over a CS told not to be followed by the unpleasant US (CS_{noUS}). In contrast, self-report ratings showed the expected effect of instructions, with more negative evaluations of the CS_{neg} than the CS_{noUS}.

The present research set out to explore this unexpected pattern of results. Experiment 1 replicated the unexpected effect of instructions on IAT scores when the IAT immediately followed the EC instructions. This ruled-out an alternative explanation of the IAT effect observed in De Houwer et al. (2019) in terms of actual pairings that occurred after the EC instructions and before the IAT. Experiments 2 and 3 further replicated the IAT effect and did not reveal differential effects on different types of IAT measures such as the personalized IAT (Experiment 2), and the recoding-free IAT (Experiment 3). Experiment 3 also did not show an effect of whether instructions referred to a contrast CS that was never followed by the aversive sound (CS_{noUS}) or that was always followed by a pleasant sound (CS_{pos}). Thus, the three experiments suggested that the unexpected effect on the IAT was robust and did not depend on manipulations of the IAT structure or instructions about the contrast CS.

Building on the results of three pilot studies, we hypothesized that the IAT effect might reflect the labelling of the CSs, such that participants readily refer to the CS_{neg} in positive terms (e.g., as the stimulus that goes together with the "crucial" or "arousing" sound), which facilitates categorization of the CS_{neg} in positive terms in the IAT. From this perspective, instructions that explicitly refer to the valence of the US might facilitate labelling the CS_{neg} in negative terms (e.g., as the CS that goes together with the negative sound), which might eliminate the unexpected IAT effect. Although we did not observe a significant typical IAT effect in the explicit instructions condition of Experiment 4, the unexpected effect found in prior studies was absent. Moreover, the type of instructions had a significant effect on the IAT effect, although this effect was weak and significant only in a (planned) one-sample t-test. In sum, our studies shed new light on the conditions under which the unexpected IAT effect does and does not arise.

The fact that instructions about the US were crucial in obtaining the unexpected IAT effect fits well with prior research on the malleability of IAT scores. These prior studies showed that labelling of IAT categories is an important determinant of IAT scores. For instance, whether Michael Jordan is evaluated as positive or negative in an IAT depends on whether that person is labelled as being a Black person or as being an athlete (Mitchell et al., 2003). Likewise, in our studies, IAT scores might have depended on whether participants labelled the CS_{neg} in negative terms (e.g., as the stimulus that co-occurs with the negative sound) or in positive terms (e.g., as the stimulus that co-occurs with the "crucial" or "arousing" sound). This labelling of the CS_{neg} might in its turn have depended on whether and how the US was labelled by the experimenter (e.g., as a "negative" sound). Note that this would be a very subtle malleability effect. Whereas previous studies manipulated the way in which an IAT category was labelled (e.g., label Michael Jordan as a Black person or as an athlete), we only varied the way in which stimuli related to an IAT category were la-

belled (i.e., whether the sound that would go together with the CS_{neg} was described as "negative").

Although our results make sense from the perspective of the malleability of IAT scores, it is important to realize that they provide only weak evidence for such malleability. The only direct evidence comes from the significant effect of instruction type on the IAT effects in Experiment 4. However, this effect was small and Bayes Factors were inconclusive. Instruction type might have had only a small effect in Experiment 4 because instructions were presented only after participants had listened to the two sounds. The direct experience with the USs at the beginning of the session might have affected CSs labelling and could have counteracted the effect of instructions. Future studies might thus replicate Experiment 4 while presenting the US instructions before the USs are presented or might eliminate the presentation of the USs altogether. Also note that whereas CS_{neg} was told to be followed by the unpleasant US "in eight trials out of ten trials", CS_{pos} was told to be "always followed the pleasant sound". This inconsistency might have somehow encouraged participants to ascribe greater salience to the aversive US (e.g., "The US that is more exceptional, or more difficult to predict given the relevant CS") irrespective of the explicit instructions about the USs. Thus, a direct follow-up study could compare IAT effects in a condition in which both USs are matched in terms of statistical contingency.

It would also be interesting to examine the role of US labelling in paradigms that produce strong typical IAT effects of EC instructions (i.e., liking of CS_{pos} over CS_{neg}). In the present studies, we started from a procedure that produces unexpected IAT effects of EC instructions and tried to eliminate this unexpected effect. In a similar vein, future studies could examine whether typical IAT effects of EC instructions can be reduced or even eliminated when USs are not labelled as positive or negative stimuli. It is indeed striking to see that most if not all studies that documented effects of EC instructions on IAT scores used instructions in which USs were labelled explicitly as positive or negative (but see Footnote 5).

If the explicit labelling of the USs is found to systematically influence (typical and unexpected) effects of EC instructions on IAT scores, the next step would be to examine whether similar US labelling effects can be found also on other indices of automatic evaluations, such as evaluative priming effects (e.g., Fazio et al., 1986). If this is the case, it would strengthen the conclusion that US labelling influences automatic evaluation of the CSs. If the effects are specific to the IAT task, it might point at non-evaluative influences on IAT performance, more specifically, the impact of US labels on the strategies that participants use to optimize IAT performance (e.g., Meissner et al., 2019; Rothermund & Wentura, 2004). Because it is currently not clear whether the IAT effects that we observed reflect automatic evaluation or non-evaluative influences, we refrain from drawing conclusions about the mental processes involved in EC and automatic stimulus evaluation.

In sum, across four experiments we found that instructions about upcoming pairings of neutral stimuli with an unpleasant (fearful sound) US can lead to an unexpected (reversed) effect on the IAT, confirming previous findings by De Houwer et al. (2019). Experiments 1-3 ruled-out several potential explanations for this effect (e.g., related to the moment of IAT administration, IAT type, or the instructions about the control stimulus). In Experiment 4, we did succeed in eliminating the unexpected IAT effect by explicitly labelling the aversive US as a negative stimulus. Although this finding is in line with previous studies showing the malleability of IAT effects, future studies are required to verify whether US labelling significantly moderates unexpected and standard IAT effects of EC instructions and whether these US labelling effects also arise for other measures of automatic evaluation.

Contributions

SM, PVD and JDH developed the research design, SM prepared the experimental materials, SM and PVD performed the analysis, SM and PVD drafted the initial manuscript, SM, PVD and JDH revised the manuscript.

Competing Interests

The authors declare that they have no conflict of interest.

Data Accessibility Statement

For all the experiments, we pre-registered the materials, sampling plan, exclusion rules, analysis plans, and hypotheses on the Open Science Framework. We report all data exclusions, manipulations, measures, and how we determined our sample sizes. The materials, data and analysis scripts of all experiments are available at osf.io/f6u2n/.

Submitted: November 02, 2020 PST, Accepted: December 11, 2020 PST

REFERENCES

De Houwer, J. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass*, *8*(7), 342–353. https://doi.org/10.1111/spc3.12111

Baeyens, F., Hermans, D., & Eelen, P. (1993). The role of CS-US contingency in human evaluative conditioning. *Behaviour Research and Therapy*, *31*(8), 731–737. https://doi.org/10.1016/0005-7967(93)90003-d

Bar-Anan, Y., Nosek, B. A., & Vianello, M. (2009). The sorting paired features task: A measure of association strengths. *Experimental Psychology*, *56*(5), 329–343. https://doi.org/10.1027/1618-3169.56.5.329

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, *61*(1), 27–41.

Boschen, M. J., Parker, I., & Neumann, D. L. (2007). Changes in implicit associations do not occur simultaneously to Pavlovian conditioning of physiological anxiety responses. *Journal of Anxiety Disorders*, *21*(6), 788–803. https://doi.org/10.1016/j.janxdis.2006.11.007

De Houwer, J. (2006). Using the Implicit Association Test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. *Learning and Motivation*, *37*(2), 176–187. https://doi.org/10.1016/j.lmot.2005.12.002

De Houwer, J. (2007). A conceptual and theoretical analysis of evaluative conditioning. *The Spanish Journal of Psychology*, *10*(2), 230–241. https://doi.org/10.1017/s1138741600006491

De Houwer, J., Mattavelli, S., & Van Dessel, P. (2019). Dissociations between learning phenomena do not necessitate multiple learning processes: Mere instructions about upcoming stimulus presentations differentially influence liking and expectancy. *Journal of Cognition*, *2*(1). https://doi.org/10.5334/joc.59

De Houwer, J., Van Dessel, P., & Moran, T. (2020). Attitudes beyond associations: On the role of propositional representations in stimulus evaluation. *Advances in Experimental Social Psychology*, *61*, 127–183.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*(2), 229–238. https://doi.org/10.1037/0022-3514.50.2.229

Field, A. P., & Lawson, J. (2003). Fear information and the development of fears during childhood: Effects on implicit fear responses and behavioural avoidance. *Behaviour Research and Therapy*, *41*(11), 1277–1293. https://doi.org/10.1016/s0005-7967(03)00034-2

Gast, A., & De Houwer, J. (2013). The influence of extinction and counterconditioning instructions on evaluative conditioning effects. *Learning and Motivation*, *44*(4), 312–325. https://doi.org/10.1016/j.lmot.2013.03.003

Gawronski, B., & Mitchell, D. G. V. (2014). Simultaneous conditioning of valence and arousal. *Cognition and Emotion*, *28*(4), 577–595. https://doi.org/10.1080/02699931.2013.843506

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197–216. https://doi.org/10.1037/0022-3514.85.2.197

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology*, *90*(1), 1–20.

Halbeisen, G., & Walther, E. (2016). Evaluative conditioning is sensitive to the encoding of CS-US contingencies. *Social Cognition*, *34*(5), 462–479. https://doi.org/10.1521/soco.2016.34.5.462

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin*, *136*(3), 390–421. https://doi.org/10.1037/a0018916

Kattner, F. (2014). Reconsidering the (in)sensitivity of evaluative conditioning to reinforcement density and CS-US contingency. *Learning and Motivation*, *45*, 15–29. https://doi.org/10.1016/j.lmot.2013.09.002

Kurdi, B., & Banaji, M. R. (2017). Repeated evaluative pairings and evaluative statements: How effectively do they shift implicit attitudes? *Journal of Experimental Psychology: General*, *146*(2), 194–213. https://doi.org/10.1037/xge0000239

Mann, T. C., Kurdi, B., & Banaji, M. R. (2020). How effectively can implicit evaluations be updated? Using evaluative statements after aversive repeated evaluative pairings. *Journal of Experimental Psychology: General*, *149*(6), 1169–1192. https://doi.org/10.1037/xge0000701

Meissner, F., Grigutsch, L. A., Koranyi, N., Müller, F., & Rothermund, K. (2019). Predicting Behavior With Implicit Measures: Disillusioning Findings, Reasonable Explanations, and Sophisticated Solutions. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02483

Mitchell, J. P., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluation. *Journal of Experimental Psychology: General*, *132*(3), 455–469. https://doi.org/10.1037/0096-3445.132.3.455

Moors, A., & De Houwer, J. (2006). Automaticity: A Theoretical and Conceptual Analysis. *Psychological Bulletin*, *132*(2), 297–326. https://doi.org/10.1037/0033-2909.132.2.297

Moran, T., & Bar-Anan, Y. (2013). The effect of object-valence relations on automatic evaluation. *Cognition & Emotion*, *27*(4), 743–752. https://doi.org/10.1080/02699931.2012.732040

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2015). Processing goals moderate the effect of co-occurrence on automatic evaluation. *Journal of Experimental Social Psychology*, *60*, 157–162. https://doi.org/10.1016/j.jesp.2015.05.009

Raes, A. K., De Houwer, J., De Schryver, M., Brass, M., & Kalisch, R. (2014). Do CS-US pairings actually matter? A within-subject comparison of instructed fear conditioning with and without actual CS-US pairings. *PLoS One*, *9*(1), e84888. https://doi.org/10.1371/journal.pone.0084888

Rescorla, R. A. (1966). Predictability and number of pairings in Pavlovian fear conditioning. *Psychonomic Science*, *4*(11), 383–384. https://doi.org/10.3758/bf03342350

Rothermund, K., Teige-Mocigemba, S., Gast, A., & Wentura, D. (2006). Minimizing the Influence of Recoding in the IAT: The IAT-RF. *Quarterly Journal of Experimental Psychology*.

Rothermund, K., & Wentura, D. (2001). Figure-Ground Asymmetries in the Implicit Association Test (IAT). *Zeitschrift für experimentelle Psychologie*, *48*(2), 94–106. https://doi.org/10.1026//0949-3946.48.2.94

Rothermund, K., & Wentura, D. (2004). Underlying Processes in the Implicit Association Test: Dissociating Salience From Associations. *Journal of Experimental Psychology: General*, *133*(2), 139–165. https://doi.org/10.1037/0096-3445.133.2.139

SUPPLEMENTARY MATERIALS

Peer Review History

Download: https://collabra.scholasticahq.com/article/18733-why-does-the-iat-reveal-a-preference-for-stimuli-said-to-be-paired-with-an-unpleasant-sound-stalking-the-unexpected/attachment/49489.docx?auth_token=fGZ26lDsJKFhMlPzmf00